Run-to-run retention time alignment improves peak picking in Skyline

Nicholas Shulman[1]; Deanna L. Plubell[1]; Brendan MacLean[1]; Michael J. MacCoss[1]
[1]University of Washington, Seattle, WA

Introduction

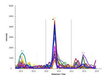
Skyline is an open-source Windows application for analyzing mass spectrometry results. Skyline has automated peak picking algorithms and allows the user to graphically see results and manually adjust peak boundaries. Traditionally, Skyline's peak scoring algorithms have focused on a single data file at a time, and incorrect peaks would often be chosen in files where the analyte was difficult to detect, despite the analyte being found in other similar runs. This shortcoming often requires using peak boundaries from other tools, or manually adjusting thousands of peak boundaries.
The next version of Skyline will make use of run-to-run alignment to improve peak detection and scoring in the runs where the analyte was difficult to detect.

Methods

A calibration curve of cerebrospinal fluid diluted into chicken serum was constructed with 100%, 70%, 50%, 30%, 10%, 5%, 1%, 0% CSF. Both CSF and SILAC-labeled and -unlabeled HeLa cell lysates was digested with a PAC bead-based protocol. HeLa digests were combined to produce 100%, 70%, 50%, 25%, 10%, 5%, 1%, 0.5%, and 0% unlabeled samples. Peptides were separated on a Vanquish Neo UHPLC, and analyzed by DIA on an Orbitrap Astral and either DIA or PRM on an Orbitrap Lumos run in development mode. DIA data was searched with either EncyclopeDIA or Chymeris, and data extracted and analyzed in Skyline.

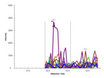
Problem

Skyline has a limitation which is particularly noticeable in calibration curve experiment where a peptide might be easily detected in the higher concentration samples but difficult or impossible to detect in the lower concentration samples.

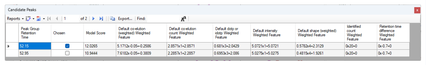


In this high-concentration sample, Skyline chooses this peak which has a strong dot product versus the library spectrum

Feature	Weighting	Score in good replicate (95% concentration)	Score in bad replicate (1% concentration)
Co-elution	0.05	0	5.1712
Co-elution count	1	4	2.8571
Library dot product	3	0.8915	0.681
Intensity	1	5.182	5.072
Shape	4	0.9669	0.5782
Total		15.7241	12.54346

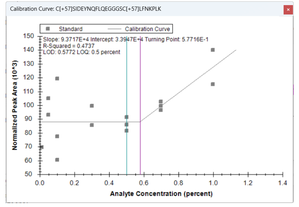


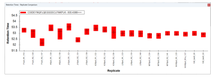
In this lower-concentration sample, none of the candidate peaks score well and Skyline ends up choosing the incorrect peak



Skyline chooses the best-looking peak in each of the samples. In the lower concentration samples Skyline typically chooses an incorrect peak which has an amount of signal which is much different from the correct value.

For the samples below the limit of detection, Skyline has chosen an incorrect peak whose area does not reflect the analyte concentration

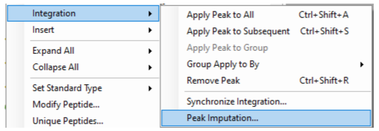




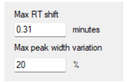
The inter-sample retention time variation is greater in the lower concentration samples because the incorrect peak was chosen

Solution

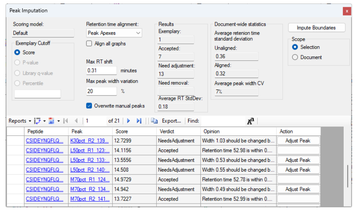
A future version of Skyline will offer a "Peak Imputation" menu item which will tell Skyline to use the peak boundaries from the better-looking samples to choose again the peaks in the samples whose retention time is very different.

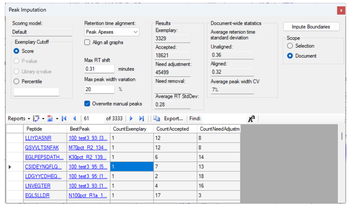


This feature requires that the user specify the allowable retention time shift and peak width variation from the best replicates before Skyline will adjust the peak boundaries of the lower-scoring replicates. In theory, these values could be determined by Skyline looking at the variation across the entire dataset, but exactly how that should happen has not been determined.

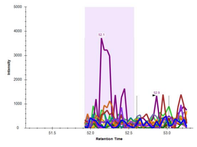


There is a details view which shows one row per replicate along with a description of what should happen and a button to change the peak boundaries in that one replicate.

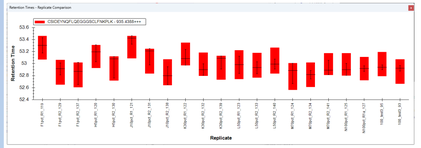




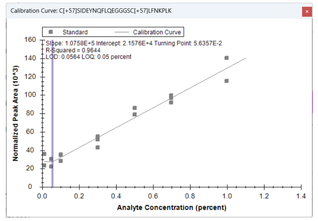
There is also a document-wide view which shows all the peptides and the number of peaks that will be adjusted.



After peak imputation, Skyline adjusts the peak boundaries of the lower-scoring replicates so that their widths and retention time more closely match the higher-scoring replicates



There is less variation in retention time between replicates.



The calibration curve looks much better. The points below the limit of detection have lower intensities than the points above that limit.

Retention time alignment

Before alignment

After alignment

By performing retention time alignment, the inter-replicate variation is reduced, and the outliers become more prominent.
Retention time alignment allows specifying a smaller "Maximum Retention Time Shift" for identifying peaks that need to be adjusted.

Aligning to a common reference

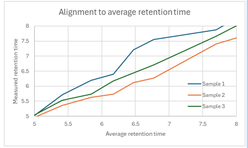
Often retention time alignment is performed between pairs of replicates. This can be complicated and might result in numerous pairwise alignments needing to be performed.

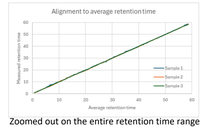
	Sample 1	Sample 2	Sample 3
SADGSPALK	6.2	5.63	5.73
GGSISGGGYGSGGGK	6.36	5.55	6.06
TESSGGWQNR	6.4	5.74	6.17
VDVDCCEK	7.29	6.89	6.06
VSTEVDAR	6.62	6	6.37
ESQAYYQR	7.21	6.12	6.44

Retention time of the apexes of the chosen peak in different replicates. The shaded cells are retention times which are out of order compared to where other replicates found the peptide.

The retention time of the peak apex of each of these consensus peptides is plotted against the average retention time of that peptide across all of the replicates.

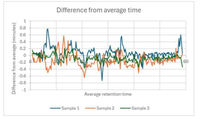
Graph zoomed in on the peptides from the above grid





Zoomed out on the entire retention time range

Plot of the difference between the retention time of each peptide in a particular replicate versus the average retention time across all replicates



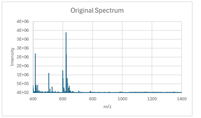
Aligning to a common reference simplifies the implementation but further investigation is needed into whether the lack of smoothness leads to incorrect peak boundary imputation.

Conflict of Interest Statement

The MacCoss Lab at the University of Washington has a sponsored research agreement with Thermo Fisher Scientific. Michael J. MacCoss is a paid consultant for Thermo Fisher Scientific. Five major mass spectrometry vendors (Agilent, Sciex, Shimadzu, ThermoFisher and Waters) provide financial support for the development and maintenance of Skyline.

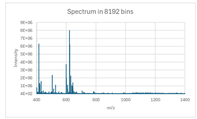
Aligning spectra between runs

It is believed that results can be further improved by looking for similar spectra between runs and using that to fine-tune the alignment.



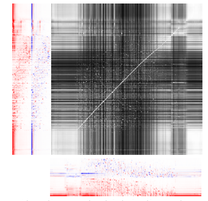
The technique for identifying similar spectra relies on first distilling the data in each spectrum down to 128 values.[1]

The values in the spectrum are first partitioned into 8192 equally spaced bins



The Haar wavelet transform is successively applied six times to the binned data reducing the number of values from 8192 to 128

Dot products are calculated between all pairs of spectra in two runs. In this graph the colored sections along the axes are heat maps representing the intensity values from the Haar wavelet transformed spectra. The grayscale chart represents the dot product values of the spectra on each axis with white points representing the most similar pairs of spectra.



There is a very clear white line along the diagonal which indicate how the retention times between the two runs should be aligned. There are also white regions closer to ends of the runs where all spectra seem to have similarity. There are heuristics that could be used to exclude those pairs from the alignment. There might be a better way to compare these spectra other than by calculating dot products.

Conclusions

These features will be available in Skyline-daily in late summer 2024.
A preview version can be installed from here:
https://proteome.gs.washington.edu/~nicksh/SpecialSkylines/PeakImputation/

References

[1]Remes PM, Yip P, MacCoss MJ. Highly Multiplex Targeted Proteomics Enabled by Real-Time Chromatographic Alignment. Anal Chem. 2020 Sep 1;92(17):11809-11817. doi: 10.1021/acs.analchem.0c02075. Epub 2020 Aug 12. PMID: 32867497; PMCID: PMC7757911.